Measurement and Evaluation in Counseling and Development
http://mec.sagepub.com/

Recalculation of the Critical Values for Lawshe's Content Validity Ratio

F. Robert Wilson, Wei Pan and Donald A. Schumsky

Measurement and Evaluation in Counseling and Development published online 14 March 2012
DOI: 10.1177/0748175612440286

The online version of this article can be found at:
http://mec.sagepub.com/content/early/2012/03/12/0748175612440286

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
AACE
ASSOCIATION FOR ASSESSMENT IN COUNSELING AND EDUCATION
Association for Assessment in Counseling and Education

Recalculation of the Critical Values for Lawshe's Content Validity Ratio

Measurement and Evaluation in Counseling and Development
XX(X) 1–14

DOI: 10.1177/0748175612440286
http://mecd.sagepub.com

F. Robert Wilson[1], Wei Pan[1], and Donald A. Schumsky[1]

Abstract

The content validity ratio (Lawshe) is one of the earliest and most widely used methods for quantifying content validity. To correct and expand the table, critical values in unit steps and at multiple alpha levels were computed. Implications for content validation are discussed.

Keywords

measurement, content validity, content validity ratio, table of critical values

Content validation rests on demonstration that the test's items are a representative sample of all items within the content domain of interest (Anastasi & Urbina, 1997; Kerlinger, 1986). Whether the researcher is evaluating the items on a test, questions in an interview, or elements of a set of accreditation standards, the items, questions, themes, or elements should all reflect the intended content of the evaluation tool (Basham & Sedlacek, 2009). Fitzpatrick (1983) described six distinct views of content validity, including four that focus on the test items—clarity of the content domain, relevance of test content to the content domain, sampling adequacy of the test content, and the technical quality of the test items. Two others focused on the test responder—sampling adequacy of test responses and relevance of test responses to a behavioral universe. Spanning the breadth of views identified by Fitzpatrick, a centrist definition for content validity might be phrased,

> Content validity of a measurement instrument for a theoretical construct reflects the degree to which the measurement instrument spans the domain of the construct's theoretical definition; it is the extent to which a measurement instrument captures the different facets of a construct. (Rungtusanatham, 1998, p. 11)

However, many assessment tools are developed for more practical reasons. An assessment tool's content validity is crucial when its scores are used as evidence in making decisions affecting an examinee's access to an educational or occupational opportunity, retention, or promotion. Lawshe (1975), an industrial–organizational psychologist with expertise in job performance assessment, speaking about the late 1960s and early 1970s, noted that "civil rights legislation, the attendant actions of compliance agencies, and a few landmark court cases have provided the impetus for the extension of the application of

[1]University of Cincinnati, Cincinnati, OH, USA

Corresponding Author:
F. Robert Wilson, University of Cincinnati, 445 TC/Dyer, Cincinnati, 45221-0002, OH, USA
Email: f.robert.wilson@uc.edu

content validity from academic achievement testing to personnel testing in business and industry" (p. 563). Decrying the lack of literature on content validity for employment assessment, he argued, "until professionals reach [consensus] regarding what constitutes acceptable evidence of content validity, there is a serious risk that the courts and enforcement agencies will play the major determining role" (p. 563). In an effort to advance the scholarship of assessment in employment settings he proposed, "Content validity is the extent to which communality or overlap exists between (a) performance on the test under investigation and (b) ability to function in the defined job performance domain" (p. 566).

Content validity is established by design and evaluated by rational analysis of test content by qualified experts in the domain of content to be assessed (Allen & Yen, 2002). To establish content validity, assessment designers follow a multistep process that includes defining the content domain and its facets, defining the level of difficulty or abstraction for the items, developing a pool of prospective items for each defined facet of the content domain, and determining domain relevant sampling ratios (Anastasi & Urbina, 1997). Some test authors might argue that if correct process was strictly followed, a content valid instrument must surely follow. Best practices in test development, however, use postdevelopment assessment of the instrument, based on a rational analysis by experts, of the representativeness (the extent to which each item within each facet of the domain of content reflects the facet's content definition) and sampling adequacy (the extent to which all aspects of a facet are adequately covered by items; Reynolds, Livingston, & Willson, 2009). This process was aided greatly by the development of methods for quantification of the expert's judgments, the first of which was the content validity ratio (CVR; Lawshe, 1975).

Lawshe introduced his method for quantifying content validity at the small, invitational Content Validity Conference held at Bowling Green University in October 1974 (Guion, 1974). Subsequently, according to Guion, Lawshe's colleague, Lowell Schipper, calculated critical values for a selection of subject matter expert (SME) sample sizes to permit significance testing. As will be shown, Lawshe's statistic has filled a need, becoming an internationally recognized method for establishing the content validity of instrumentation across many disciplines. Developed at a time when statistical analysis in the social sciences relied on submitting data recorded on Hollerith punch cards into mainframe computers, Lawshe's item-level CVR, and its multi-item summary statistic, the Content Validity Index, when coupled with Schipper's table of critical values, provided an easy-to-compute method for quantification and significance testing in studies of content validity.

Unfortunately, whether due to a calculation error, a typographical error, or a typesetter's error, Schipper's table of critical values appears to contain an anomaly. Although distributions of critical values are typically monotonic, Schipper's table contains a discontinuity (noted by Stelly, 2006). Moreover, there is apparently no record of how Schipper computed the set of critical values Lawshe published. The purpose of this study, therefore, was to identify how Schipper's values were computed and then to recompute the table of critical values to correct the discontinuity.

Lawshe's Content Validity Methodology

Following established methodology, Lawshe's approach called for the assembly of a set of SMEs who rated each of an instrument's items on a 3-point scale: (a) "essential," (b) "useful, but not essential," and (c) "not necessary." His statistic, the content validity ratio or CVR, was a linear transformation of the ratio of the number of SMEs judging an item to be "essential" to the total number of SMEs in the panel. Specifically,

$$\mathrm{CVR} = \frac{n_{\mathrm{e}} - (N/2)}{N/2},$$

where n_{e} is the number of SMEs indicating that the item is "essential," and N is the total number of SMEs in the panel.

When all SMEs rate the item as being "essential," the value of CVR will compute to be 1; when the number rating the item as "essential" is more than half but less than all, the value of CVR will be between 0 and 1; and when less than half of the SMEs rate the item as "essential," the value of CVR will be negative. Although this statistic is no more than a linear transformation of the proportion of SMEs judging the item as "essential," Lawshe's true contribution was in providing a table of critical values, which he attributed to his colleague Lowell Schipper, for determining whether the SMEs' judgments exceeded chance expectation at a one-tailed alpha level of .05.

Compared with alternative methods for quantifying content validity judgments, the Lawshe method is straightforward and user-friendly, requiring only simple computations and providing a table for determining a critical cutoff value. Alternative methods such as Cohen's kappa (κ; Cohen, 1960), the Tinsley–Weiss *T* index (Tinsley & Weiss, 1975), James, Demaree, and Wolf's (1993) r_{WG} and $r_{WG(J)}$ indexes, and Lindell, Brandt, and Whitney's (1999) $r^*_{WG(J)}$ indexes are more computationally complex than Lawshe's CVR and focus on interrater agreement in general rather than on the specific issue of agreement that an item is "essential" (Lindell & Brandt, 1999).

Critical Acceptance of Lawshe's Methods

Since its introduction in 1975, critical acceptance of Lawshe's CVR methodology has grown. The popularity of the Lawshe approach in scale development for health and education sciences is demonstrated by the number of articles published making reference to the CVR and by the wide ranging studies in which it has been used. An electronic search of the *Summon* electronic database revealed 94 articles containing the phrase, "content validity ratio" of which 51 were published in the past 5 years.

Prevention and health promotion specialists have used Lawshe's CVR to develop scales for assessing child-rearing knowledge and practices for women with epilepsy (Saramma & Thomas, 2010), a belief-based physical activity questionnaire for diabetic patients (Ghazanfari, Niknami, Ghofranipour, Hajizadeh, & Montazeri, 2010), a checklist for performing content analysis on patient education course syllabi (Gail-Hinckley Heitzer, McKenzie, Amschler, & Bock, 2009), and for assessing whether generic quality of life scales were free of content related to physical function (Hall, Krahn, Horner-Johnson, & Lamb, 2011). In the field of mental health and rehabilitation, researchers developed scales for assessing feelings of competence among children with attention-deficit/hyperactivity disorder (ADHD; Hanc & Brzezinska, 2009), satisfaction with treatment for sexual dysfunction (Corty, Althof, & Wieder, 2011), and psychotherapist countertransference (Hayes, 2004) using CVR methodology to assess content validity. In a novel study, cross-cultural researchers used the CVR to determine the cultural relevance of items drawn from the Indiana Job Satisfaction Scale (IJSS) thereby producing a Chinese version of the IJSS for use in vocational rehabilitation programs for individuals with mental retardation in China (Tsang & Wong, 2005).

Medical and nursing assessment specialists have relied on the Lawshe approach for developing an adult intubation procedural checklist (Stausmire, 2011), a quality-of-life index for AIDS patients in Uganda (Namisango, Katabira, Karamagi, & Baguma, 2007), a system for auditing nursing care plans (Bjorvell, Thorell-Ekstrand, & Wredling, 2000), a low-literacy assessment of patient knowledge regarding chronic obstructive pulmonary disease (Maples, Franks, Ray, Stevens, & Wallace, 2010), a survey to assess Medicaid recipients' understanding of the postpartum tubal sterilization process (Zite & Wallace, 2007), and a Swedish language version of the Problem Areas in Diabetes Scale (Amsberg, Wredling, Lins, Adamson, & Johansson, 2008).

In the field of education, the content validity of a scale for evaluating team-designed material development manuals (Erdem, 2009) and an affective response to literature scale (Fischer & Fischer, 2007) was established by

SMEs working according to Lawshe's methods. Training specialists have used the CVR to assess job relatedness of the content of a job training program (Ford & Wroten, 2006) and the job relatedness of an assessment of posttraining job knowledge (Distefano, Pryer, & Craig, 2006).

Organizational developers and management specialists have used Lawshe's content validity approach to assess the impact of the Deming model for quality management (Collard, 1992) and to define and measure *servant leadership* behavior (Sendjaya, Sarros, & Santora, 2008). A series of studies based on applications of the enterprise resource planning model in Asian business markets, has used the CVR to develop performance indicators or critical success factors (J. Huang, Zhao, & Li, 2007; S.-M. Huang, Hung, Chen, & Ku, 2004; Wei, 2008; Yu, Ng, Chang, Chang, & Yen, 2011). Drossos and Fouskas (2010) used the CVR to assess the content validity of a tool developed to measure industry perceptions of the competitiveness of market environments and their own competitive responses.

Market research has also embraced the Lawshe method for assessing content validity. Tools for assessing consumer adaption to or adoption of broadband (Choudrie, Dwivedi, & Brinkman, 2006), Internet stock trading (Hung, Huang, & Yen, 2004), and airport self-service check-in kiosks (Chang & Yang, 2008a) were developed using CVR methodology. The CVR was also used in developing criteria to segment a customer base (Tai, 2011; Tai & Ho, 2010), assess brand personality appeal (Henard, Freling, & Crosno, 2011), and assess passenger repurchase motivation (Chang & Yang, 2008b). Concern over issues of internet security prompted the development of tools for assessing perceived functional and relational value of information sharing services (Tai, 2011) and for assessing privacy concerns and levels of information exchange for e-services on the Internet (Dinev & Hart, 2006), both of which were CVR-supported research tools.

In the field of personnel psychology, Lawshe's methodology has been used in the development of a situational interview to predict service representative applicants' future job performance (Flint & Haley, 2008), a structured behavioral interview to hire private security personnel (Moscoso & Selgado, 2001), a job performance rating criterion (Distefano, Pryer, & Erffmeyer, 2006), and job termination criteria for assessing mentally ill workers (Mak, Tsang, & Cheung, 2006). Mathews, Smith, Hussey, and Plack (2010) used the CVR to develop an assessment tool to measure participants' perceptions of the roles, practices, education, and preferred relationship of physical therapists and physical therapist assistants. Finally, Lawshe's CVR was also used to develop tools for assessing critical factors related to Taiwanese expatriates' foreign post selection and overseas performance (Cheng & Lin, 2009).

Despite its competitors (e.g., Cohen's κ, 1960; the Tinsley–Weiss *T*-Index, Tinsley & Weiss, 1975; James et al.'s r_{WG} and $r_{WG(J)}$ indexes, 1993; and Lindell et al.'s $r^*_{WG(J)}$ index, 1999), the Lawshe method has been endorsed in texts on personnel management (Lindell & Brandt, 1999) and endorsed for use in nursing research (Polit & Beck, 2006; Polit, Beck, & Owen, 2007). Its tabled critical values have been reproduced in texts such as the Cohen and Swerdlik (2005) text on psychological testing and assessment.

Problems With Schipper's Table of Critical Values

Though Lawshe's method has received commendation and has been featured in research studies in multiple disciplines, and is even being used in defense of the content validity of high-stakes tests, it is not without criticism. The main thrust of the criticism has been directed toward three aspects of the table of critical values Lawshe provided for the CVR, a table which Lawshe acknowledges was developed by his colleague, Lowell Schipper.

Schipper's table of critical values is terse. It only provides critical values for pools of 5, 6, 7, . . ., 14, 15, 20, 25, 30, 35, and 40 SMEs and this only for one alpha level. Although interpolation of missing data points with a

linear function can be accomplished easily, a scatter plot of Schipper's table reveals that the critical values are curvilinear making accurate interpolation problematic.

A careful examination of the critical values also reveals an anomaly. The critical value for the CVR increases monotonically from the case of 40 SMEs ($CVR_{critical}$ = .29) to the case of 9 SMEs ($CVR_{critical}$ = .78) only to unexpectedly drop at the case of 8 SMEs ($CVR_{critical}$ = .75) before hitting its ceiling value at the case of 7 SMEs ($CVR_{critical}$ = .99). When Cohen and Swerdlick (2005) reproduced Schipper's table in their assessment text, they did not comment on this apparent anomaly. When Wallace, Gregory, Parham, and Baldridge (2003) used the CVR method with nine SMEs to develop and validate family residency recruitment questionnaires, they used a $CVR_{critical}$ of .75. Whether using a $CVR_{critical}$ of .75 at $N = 9$ was an error on their part in reading Schipper's table or an attempt to adjust for the apparent anomaly at $N = 8$ is unknown. On reviewing Wallace et al.'s (2003) work, Stelly (2006) observed, "it is possible that the authors reversed the minimum CVRs for 8 and 9 panelists to correct what they perceived to be an error in the original table" (p. 6). The anomaly may also be a function of something as simple as a typographical error which escaped proofreading, or perhaps a typesetter's error given the fact that in the 1970s, many journals used hand-set type for tables, if not for the whole of the journal.

But the most unsettling problem is that the statistical distribution underlying Lawshe's table is not specified. In his defining article, Lawshe (1975) stated that the table of critical values for the CVR was calculated by his friend, Lowell Schipper. Unfortunately, he did not describe the basis on which these values were calculated. Lawshe had introduced the CVR at the 1974 Content Validity Conference, a small invitational conference sponsored by the Society for Industrial and Organizational Psychology (SIOP). Another SIOP member, Robert M. Guion (1974) wrote,

> C. H. Lawshe presented a scheme for classifying content validity problems and a "Content Validity Ratio" by which the relevance of a test item or a total test score might be scaled. (Lowell Schipper has subsequently related the CVR to chi square, permitting significance testing). (p. 18)

Apparently not having had access to Guion's review of this conference, Lindell and Brandt (1999) and Stelly (2006) speculated that the critical values were associated with the binomial distribution.

Purpose of This Investigation

Since the Lawshe method is being used to produce knowledge for diverse disciplines and its possibly flawed tabled values are being disseminated in print and electronic media, correction of the apparent errors in Lawshe's (1975) presentation of Schipper's table and extension of the range of tabled values are warranted. The purpose for this study is therefore to explore the CVR's underlying distribution and to correct and expand the range of its tabled critical values.

Do Schipper's Critical Values Map to the Binomial Distribution?

Both Lindell and Brandt (1999) and Stelly (2006) speculated that Schipper's critical values were associated with the binomial distribution, a more precise hypothesis than Guion's (1974) report of Schipper relating the CVR to chi square. To evaluate the proposition that Schipper's table of critical values for the CVR was based on the binomial distribution, two approaches were taken: (a) an examination of the cumulative probabilities for sets of independent Bernoulli trials and (b) an examination of the normal approximation for the binomial distribution.

Discrete Binomial Probabilities

To determine whether Schipper based his table of critical values on discrete binomial

probabilities, the cumulative probabilities for sets of independent Bernoulli trials were calculated. Although we expected that this approach would not yield a monotonic progression of values, it seemed important to test this approach given Stelly's (2006) advocacy for using exact probabilities.

A key parameter in these calculations is the value for p, the probability for any given trial of achieving success. The conventional way of construing the problem would be to view Lawshe's rating scale as a trichotomy, with the three outcomes being (a) "essential," (b) "useful, but not essential," and (c) "not necessary." From this point of view, the parameter, p would be ⅓. However, Lawshe construed the scale as a dichotomy, with the two outcomes being (a) "essential" and (b) "not essential" (with "useful, but not essential" and "not necessary" being combined as the second category) yielding a value for p of ½. For this exploration, both approaches were tried.

For each approach (i.e., dichotomous, trichotomous), a table of critical values based on the discrete binomial was computed using the Microsoft Excel function:

$$n_{\text{critical}} = \text{CRITBINOM}(N, p, 1-\alpha),$$

where n_{critical} is the smallest value for n (the number of SMEs judging the item as "essential") for which the cumulative binomial distribution is greater than or equal to a criterion value $1 - \alpha$, N is the number of Bernoulli trials (the number of SMEs in the pool), and p is the probability of success on each trial. Since CRITBINOM returns the smallest value for n_e (the number of SMEs judging the item as "essential"), CRITBINOM's output was converted to a value of $\text{CVR}_{\text{critical}}$ according to Lawshe's CVR formula:

$$\text{CVR}_{\text{critical}} = \frac{n_{\text{critical}} - (N/2)}{(N/2)}.$$

To obtain a complete table of values, we computed $\text{CVR}_{\text{critical}}$ for each N from 5 through 40 in unit steps. We also expanded the table by considering the traditional range of values for alpha. For each alpha level, the significance of difference between Schipper's critical values and those computed using CRITBINOM for each alpha level was tested using the nonparametric Wilcoxon signed-rank for dependent samples to determine for which, if any, of the alpha levels did the computed $\text{CVR}_{\text{critical}}$ values differ from those attributed to Schipper. Because Schipper's values achieve a ceiling value of $\text{CVR}_{\text{critical}} = .99$ at a pool size of $N = 7$, only the calculated values in the range of $N = 7, \ldots, 40$ were tested for departure from Schipper's values.

Discrete binomial construing SME ratings as a trichotomy. Examination of the proposition that Lawshe's rating scale should be treated as a trichotomy rather than Lawshe's favored dichotomy produced a poor fit to Schipper's critical values for the CVR. With the probability of success set at $p = ⅓$, for each criterion value for alpha, the distribution of binomial probabilities yielded a pronounced, jagged or "saw-toothed" pattern. The best fit as evidenced by the mean absolute deviation from Schipper's values, was found to be at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed). The mean absolute departure of the calculated values for $\text{CVR}_{\text{critical}}$ from Schipper's critical values ranged from a minimum difference of .09 at $\alpha = .001$, one-tailed (or $\alpha = .002$, two-tailed) to a maximum difference of .56 at $\alpha = .10$, one-tailed (or $\alpha = .20$, two-tailed). The Wilcoxon signed-rank test revealed that at all but one of the tested alpha levels, the difference between the calculated values for $\text{CVR}_{\text{critical}}$ departed significantly from those provided by Schipper's values ($p < .01$ for all tests). The only distribution of $\text{CVR}_{\text{critical}}$ computed using CRITBINOM with $p = ⅓$ that was sufficiently close in value to those supplied by Schipper to be considered interchangeable with his table of minimum values was at an extreme alpha level, $\alpha = .0005$, one-tailed ($\alpha = .001$, two-tailed). These results are presented in Table 1.

Discrete binomial construing SME ratings as a dichotomy. With the probability of success set at $p = ½$, for each criterion value for alpha, the distribution of binomial probabilities yielded a less pronounced "saw-toothed" pattern. The mean absolute departure of the calculated values for $\text{CVR}_{\text{critical}}$ from Schipper's critical

Table 1. Comparison of Schipper's $CVR_{critical}$ With Three Recalculations Based on the Binomial Distribution at Eight Levels of Significance

Level of Significance		Discrete Binomial								Normal Approximation to Discrete Binomial			
		Trichotomous Rating				Dichotomous Rating							
			Wilcoxon Signed-Rank Test				Wilcoxon Signed-Rank Test				Wilcoxon Signed-Rank Test		
One-Tailed Test	Two-Tailed Test	Mean Absolute Difference	*N*	*T*	*p*	Mean Absolute Difference	*N*	*T*	*p*	Mean Absolute Difference	*N*	*T*	*p*
.1	.2	.56	14	0	<.01	.20	13	14	<.05	.20	14	0	<.01
.05	.1	.47	14	0	<.01	.10	14	0	<.01	.10	14	0	<.01
.025	.05	.38	14	0	<.01	.05	13	29.5	*ns*	.04	14	40	*ns*
.01	.02	.30	14	0	<.01	.09	13	14	<.05	.09	14	11	<.01
.005	.01	.21	14	0	<.01	.12	13	1	<.01	.15	14	1	<.01
.0025	.005	.14	13	0	<.01	.17	14	1	<.01	.20	14	0	<.01
.001	.002	.09	13	14	<.01	.22	14	0	<.01	.28	14	0	<.01
.0005	.001	.07	13	23	*ns*	.25	14	0	<.01	.33	14	0	<.01

Note: CVR = content validity ratio.

values was least at α = .05, two-tailed (or α = .025, one-tailed) and when tested using the Wilcoxon signed-rank test, the calculated values at this alpha level were found to not differ significantly from Schipper's values. At all other alpha levels, the mean difference was higher (range: .09–.22) and the departure of the calculated values from those proposed by Schipper was significant with the level of significance ranging from $p = .05$ to $p = .01$. The results of these tests are presented in Table 1.

Normal Approximation to the Binomial Distribution

Although the calculation of discrete probabilities yielded values that at α = .05, two-tailed (or α = .025, one-tailed) bracketed in "sawtooth" fashion those provided by Lawshe, they failed to be monotonic. Calculation of the normal approximation to the discrete binomial calculations would yield a monotonic curve. Assuming that $n_e \sim B(N, p)$ and assuming that $p = ½$, the normal approximation of the binomial distribution may be expressed as

$$z = \frac{n_e - Np}{\sqrt{Np(1-p)}} = \frac{n_e - (N/2)}{(\sqrt{N}/2)} \sim N(0,1).$$

According to Box, Hunter, and Hunter (1978, p. 130), for $N > 5$ the normal approximation is adequate if

$$\left| \frac{1}{N} \left(\sqrt{\frac{1-p}{p}} - \sqrt{\frac{p}{1-p}} \right) \right| < 0.3.$$

In this case, $p = ½$, so the assumption above is satisfied.

The task of the CVR is to identify items in an instrument deemed by a critical number of content experts to be "essential." This task calls for a one-tailed hypothesis test, expressed as

$$H_0 : n_e \le \frac{N}{2} \text{ versus } H_1 : n_e > \frac{N}{2},$$

for which the corresponding critical value is

$$\frac{n_{e.\alpha} - (N/2)}{(\sqrt{N}/2)} = z_\alpha,$$

where α is a prespecified significance level; or

$$n_{e.\alpha} = z_\alpha \times \frac{\sqrt{N}}{2} + \frac{N}{2}.$$

Therefore, the critical value for CVR is

$$CVR_\alpha = \frac{n_{e.\alpha} - (N/2)}{(N/2)} = \frac{z_\alpha}{\sqrt{N}}.$$

As above, to obtain a complete table of values, we computed $CVR_{critical}$ for each N from 5 through 40 in unit steps and expanded the table by considering the traditional range of values for alpha. The departure of the computed $CVR_{critical}$ values from those of Schipper's was again tested using the nonparametric Wilcoxon signed-rank for dependent samples. The recalculated table of critical values is presented in Table 1.

For each level of alpha, the distribution of binomial probabilities yielded a smooth, monotonic curve. As was noted earlier, because Schipper's values achieve a ceiling value of $CVR_{critical} = .99$ at a pool size of $N = 7$, only the calculated values in the range of $N = 7, \ldots, 40$ were analyzed. The mean absolute departure of the calculated values for $CVR_{critical}$ from Schipper's critical values ranged from a maximum difference of .28 at $\alpha = .001$, one-tailed (or $\alpha = .002$, two-tailed) to a minimum difference of .04 at $\alpha = .025$, one-tailed (or at $\alpha = .05$, two-tailed). When tested using the Wilcoxon signed-rank test, the calculated values $\alpha = .025$, one-tailed (or at $\alpha = .05$, two-tailed) were found to not differ significantly from Schipper's values. At all other alpha levels, the mean difference was higher (range: .09–.28) and the departure of the calculated values from those proposed by Schipper was significant with the level of significance of $p < .01$. For small pools of judges ($N = 5, \ldots, 10$), values computed for $CVR_{critical}$ at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed) were more liberal while at increasingly larger pools of judges ($N = 20, \ldots, 40$), the values computed for $CVR_{critical}$ at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed) were slightly more conservative. The results of these tests are presented in Table 1 and the complete table of recalculated values based on the normal approximation to the binomial is presented as Table 2. A graph illustrating Schipper's values for $CVR_{critical}$ and curves showing the values for $CVR_{critical}$ calculated by the normal approximation to the binomial with the scaling construed as a dichotomy and as discrete binomial probabilities with the scaling construed both as a trichotomy and as a dichotomy at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed) is presented in Figure 1.

Discussion

The questions raised about Schipper's table of critical values for Lawshe's CVR focus on four issues: (a) How did Schipper compute the table? (b) Was Lawshe correct in labeling Schipper's table of critical values as representing a test at $\alpha = .05$, one-tailed? (c) Why does Schipper's table contain an anomaly? (d) If errors were made, what are the likely consequences of the errors?

How Did Schipper Compute His Table of Critical Values for CVR?

It appears that Schipper did not compute discrete binomial probabilities. It appears more likely that he used the normal approximation for computing binomial probabilities to create his table. Although the curve produced by calculating the normal approximation to the binomial does not provide an exact fit to Schipper's values, the curve produced at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed) is a very close approximation. Values calculated at all other alpha levels result in larger mean absolute discrepancy. The Wilcoxon test found significant discrepancy between the calculated values and Schipper's values for all alpha levels tested except at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed).

Does Schipper's Table Provide a Test at $\alpha = .05$, One-Tailed?

It also appears that Lawshe was in error in labeling Schipper's table as providing a test for $CVR_{critical}$ at $\alpha = .05$, one-tailed. As noted above, although the curve produced by calculating the normal approximation to the binomial does not fit the full range of Schipper's data exactly, the values produced at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed) provide a very close fit. A quantitative methods specialist with 50 or more years in the profession, observed that in those early years, many

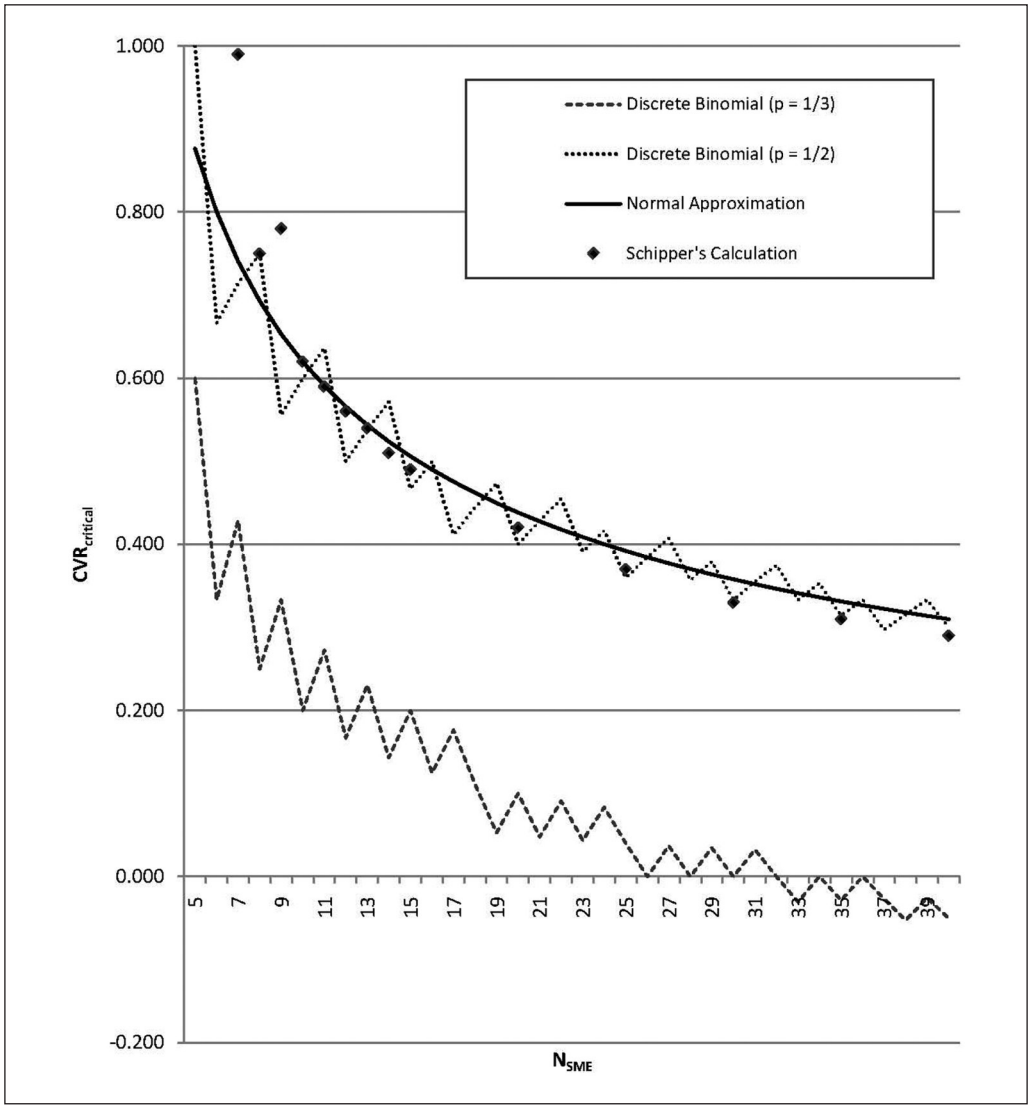

Figure 1. Comparison of Schipper's values of $CVR_{critical}$ with critical value results from three recalculations: the discrete binomial ($p = {}^{1}/_{3}$), the discrete binomial ($p = ½$), and the normal approximation to the binomial

quantitative analysts ran two-tailed tests even when the hypothesis under test was directional (D. A. Schumsky, personal communication, June 10, 2011). Perhaps Schipper, the statistician, produced a table of values at $\alpha = .05$, two-tailed out of habit and Lawshe, the theoretician and applied personnel psychologist did not realize that such was the case.

Why Does Schipper's Table Contain an Anomaly?

Although this was the question which initiated this project, it may go unanswered. We had hoped that we would be able to reproduce Schipper's values exactly (except, of course, the anomalous value). We would have then

Table 2. Critical Values for Lawshe's (1975) Content Validity Ratio ($CVR_{critical}$)

	Level of Significance for One-Tailed Test					
	.1	.05	.025	.01	.005	.001
	Level of Significance for Two-Tailed Test					
N	.2	.1	.05	.02	.01	.002
5	.573	.736	.877	.99	.99	.99
6	.523	.672	.800	.950	.99	.99
7	.485	.622	.741	.879	.974	.99
8	.453	.582	.693	.822	.911	.99
9	.427	.548	.653	.775	.859	.99
10	.405	.520	.620	.736	.815	.977
11	.387	.496	.591	.701	.777	.932
12	.370	.475	.566	.671	.744	.892
13	.356	.456	.544	.645	.714	.857
14	.343	.440	.524	.622	.688	.826
15	.331	.425	.506	.601	.665	.798
16	.321	.411	.490	.582	.644	.773
17	.311	.399	.475	.564	.625	.750
18	.302	.388	.462	.548	.607	.729
19	.294	.377	.450	.534	.591	.709
20	.287	.368	.438	.520	.576	.691
21	.280	.359	.428	.508	.562	.675
22	.273	.351	.418	.496	.549	.659
23	.267	.343	.409	.485	.537	.645
24	.262	.336	.400	.475	.526	.631
25	.256	.329	.392	.465	.515	.618
26	.251	.323	.384	.456	.505	.606
27	.247	.317	.377	.448	.496	.595
28	.242	.311	.370	.440	.487	.584
29	.238	.305	.364	.432	.478	.574
30	.234	.300	.358	.425	.470	.564
31	.230	.295	.352	.418	.463	.555
32	.227	.291	.346	.411	.455	.546
33	.223	.286	.341	.405	.448	.538
34	.220	.282	.336	.399	.442	.530
35	.217	.278	.331	.393	.435	.522
36	.214	.274	.327	.388	.429	.515
37	.211	.270	.322	.382	.423	.508
38	.208	.267	.318	.377	.418	.501
39	.205	.263	.314	.372	.412	.495
40	.203	.260	.310	.368	.407	.489

Note: Values for $CVR_{critical}$ greater than or equal to the limit value of 1.00 were set to .99.

been able to replace the anomalous value with a correct one. Several speculations have arisen. The anomaly could have arisen from a typographical error, a failure in proofreading. Another possibility, since tables in older journals were often set by hand, is that the anomaly was a result of interchanging the two lines of type containing the critical values for $N = 8$ and $N = 9$. Finally, since the values computed using the normal approximation to the binomial fit well with Schipper's values until SME pool sizes fall below 10, there may be more than a single anomaly in Schipper's table. If one presumes there is a single anomaly at $N = 9$, the anomaly could have been the result of a single calculation error. With such a small number of values to compute, Schipper may have computed the values longhand or with the aid of a calculator and may have simply made a mistake in calculating the value for $CVR_{critical}$ at $N = 9$. However, Schipper's value for $CVR_{critical}$ at $N = 7$ is also very different from that which is produced using the normal approximation to the binomial. In Schipper's table, the $CVR_{critical}$ at $N = 7$, 6, and 5 was set at a ceiling value of .99. One possibility is that these ceiling values were not calculated but were inserted, a priori, as a statement that at such small sample sizes, only perfect agreement among the SMEs that the item under scrutiny was "essential" could be accepted safely. In his 1975 article, Lawshe provided no discussion about the construction of the table. It is unfortunate that no question was raised about the anomalous value or values in Schipper's table before his death in 1984 and that there was apparently no contact between Lindell and Brandt, who published their review of methods for quantifying content validity judgments in the same year that Lawshe died (1999), to enquire about Schipper's table.

What Are the Consequences of Lawshe's and Perhaps Schipper's Apparent Errors?

The apparent mislabeling of the alpha level for Schipper's table and the presence of one or more anomalies in the table suggest that the table may lead to erroneous decisions by

the researcher. Given that Lawshe's method has been used widely, even in high-stakes testing situations, the consequences could be potentially harmful. Fortunately, both apparent errors are errors on the side of stringency:

- *Consequence of the apparent mislabeling of the table's alpha level*. Lawshe's labeling of Schipper's table as representing a test at $\alpha = .05$, one-tailed is an error in the conservative direction. Since Schipper's table of critical values appears to represent a test at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed), an item's content validity, rated at a level beyond chance expectation at a true $\alpha = .05$, one-tailed would be rejected according to Schipper's values for $CVR_{critical}$. This is an error in the conservative direction.
- *Consequence of the apparent anomaly or anomalies in Schipper's table.* Compared with the values calculated for the normal approximation to the binomial, Schipper's value for the $CVR_{critical}$ of .78 at $N = 9$ is a much more stringent criterion than the value of .653 computed at that pool size for the normal approximation to the binomial at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed). While Schipper's apparently anomalous value of .75 is closer to the normal approximation value of .693 at $N = 8$, his value of .99 at $N = 7$ is also much more stringent than the value of .741 computed for the normal approximation to the binomial.

With small pools of SMEs, a test author who used Schipper's table for setting the criterion for item inclusion would have little reason to worry about whether an item with low content validity had been included in the test. Both errors (i.e., the anomalous value or values and the apparent mislabeling of the table) lead to increasing the stringency of the criterion for item inclusion. Since Lawshe's CVR has been used to produce high-stakes employment tests, erring in the conservative direction offers greater safety from allegations that the test contained items not judged to be "essential" for job performance. Given the consequences of using an invalid test in high-stakes testing, if an error was to be made, an error in the conservative direction is the better of the two possible errors.

Conclusions

Lowell Schipper's table of critical values for Charles Lawshe's CVR, which Lawshe described as representing a test at $\alpha = .05$, one-tailed, was examined. Evidence showed that it had one or more anomalous values for $CVR_{critical}$. A review of literature failed to shed light on the method used by Schipper in calculating the table. Trial tables of critical values were computed using both discrete calculation and normal approximations to the binomial distribution. Schipper's values mapped convincingly to the normal approximation of the binomial at $\alpha = .05$, two-tailed (or $\alpha = .025$, one-tailed) suggesting that Lawshe may have mislabeled the alpha level for Schipper's table—rather than being a table of values for $\alpha = .05$, one-tailed, it is likely that it is a table of values for $\alpha = .05$, two-tailed. This finding suggests that, at small SME pool sizes, Schipper's values for $CVR_{critical}$ represent a more conservative criterion for item inclusion than may be warranted.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory* (2nd ed.). Prospect Heights, IL: Waveland Press.

Amsberg, S., Wredling, R., Lins, P., Adamson, U., & Johansson, U. (2008). The psychometric

properties of the Swedish version of the problem areas in diabetes scale (swe-PAID-20): Scale development. *International Journal of Nursing Studies*, *45*, 1319–1328. doi:10.1016/j.ijnurstu.2007.09.010

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York, NY: Prentice Hall.

Basham, A., & Sedlacek, W. E. (2009). Validity. In American Counseling Association (Ed.), *The ACA encyclopedia of counseling* (p. 557). Alexandria, VA: American Counseling Association.

Bjorvell, C., Thorell-Ekstrand, I., & Wredling, R. (2000). Development of an audit instrument for nursing care plans in the patient record. *Quality Health Care*, *9*, 6–13. doi:10.1136/qhc.9.1.6

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York, NY: Wiley.

Chang, H.-L., & Yang, C.-H. (2008a). Do airline self-service check-in kiosks meet the needs of passengers? *Tourism Management*, *29*, 980–993. doi:10.1016/j.tourman.2007.12.002

Chang, H.-L., & Yang, C.-H. (2008b). Explore airlines' brand niches through measuring passengers' repurchase motivation: An application of Rasch measurement. *Journal of Air Transport Management*, *14*, 105–112. doi:10.1016/j.jairtraman.2008.02.004

Cheng, H.-L., & Lin, C. Y. Y. (2009). Do as the large enterprises do? Expatriate selection and overseas performance in emerging markets: The case of Taiwan SMEs. *International Business Review*, *18*, 60–75. doi:10.1016/j.ibusrev.2008.12.002

Choudrie, J., Dwivedi, Y. K., & Brinkman, W. (2006). Development of a survey instrument to examine consumer adoption of broadband. *Industrial Management & Data Systems*, *106*, 700–718. doi:10.1108/02635570610666458

Cohen, R. J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, R. J., & Swerdlik, M. E. (2005). *Psychological testing and assessment*. New York, NY: McGraw-Hill.

Collard, E. F. N. (1992). *The impact of deming quality management on interdepartmental cooperation* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis.

Corty, E. W., Althof, S. E., & Wieder, M. (2011). Measuring women's satisfaction with treatment for sexual dysfunction: Development and initial validation of the Women's Inventory of Treatment Satisfaction (WITS-9). *Journal of Sexual Medicine*, *8*, 148–157. doi:10.1111/j.1743-6109.2010.01977.x

Dinev, T., & Hart, P. (2006). Privacy concerns and levels of information exchange: An empirical investigation of intended e-services use. *e-Service Journal*, *4*(3), 25–60. doi:10.2979/ESJ.2006.4.3.25

Distefano, M. K., Pryer, M. W., & Craig, S. H. (2006). Job-relatedness of a posttraining job knowledge criterion used to assess validity and test fairness. *Personnel Psychology*, *33*, 785–793.

Distefano, M. K., Pryer, M. W., & Erffmeyer, R. C. (2006). Application of content validity methods to the development of a job-related performance rating criterion. *Personnel Psychology*, *36*, 621–631.

Drossos, D. A., & Fouskas, K. G. (2010). The role of industry perceptions in competitive responses. *Industrial Management & Data Systems*, *110*, 477–494. doi:10.1108/02635571011038981

Erdem, M. (2009). Effects of learning style profile of team on quality of materials developed in collaborative learning processes. *Active Learning in Higher Education*, *10*, 154–171. doi:10.1177/1469787409104902

Fischer, R. G., & Fischer, J. M. (2007). The development of an emotional response to literature measure: The affective response to literature survey. *Alberta Journal of Educational Research*, *52*, 265–276.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement*, *7*, 3–13.

Flint, D., & Haley, L. (2008). Content oriented development of a situational interview. *The Business Review, Cambridge*, *10*(2), 21.

Ford, J. K., & Wroten, S. P. (2006). Introducing new methods for conducting training evaluation and for linking training evaluation to program redesign. *Personnel Psychology*, *37*, 651–665.

Gail-Hinckley Heitzer, J., McKenzie, J. F., Amschler, D. H., & Bock, W. (2009). A descriptive analysis of patient education courses in undergraduate and graduate health education programs. *Health Promotion Practice*, *10*, 244–253. doi:10.1177/1524839907309046

Ghazanfari, Z., Niknami, S., Ghofranipour, F., Hajizadeh, E., & Montazeri, A. (2010). Development and psychometric properties of a belief-based physical activity questionnaire for diabetic patients (PAQ-DP). *BMC Medical Research Methodology*, *10*, 104.

Guion, R. M. (1974). Content validity conference. *The Industrial-Organizational Psychologist* (Newsletter), *12*(1), 18.

Hall, T., Krahn, G. L., Horner-Johnson, W., & Lamb, G. (2011). Examining functional content in widely used health-related quality of life scales. *Rehabilitation Psychology*, *56*, 94–99. doi:10.1037/a0023054

Hanc, T., & Brzezinska, A. I. (2009). Intensity of ADHD symptoms and subjective feelings of competence in school age children. *School Psychology International*, *30*, 491–506. doi:10.1177/0143034309107068

Hayes, J. A. (2004). The inner world of the psychotherapist: A program of research on countertransference. *Psychotherapy Research*, *14*, 21–36. doi:10.1093/ptr/kph002

Henard, D. H., Freling, T. H., & Crosno, J. L. (2011). Brand personality appeal: Conceptualization and empirical validation. *Journal of the Academy of Marketing Science*, *39*, 392–406. doi:10.1007/s11747-010-0208-3

Huang, J., Zhao, C., & Li, J. (2007). An empirical study on critical success factors for electronic commerce in the Chinese publishing industry. *Frontiers of Business Research in China*, *1*, 50–66. doi:10.1007/s11782-007-0004-1

Huang, S.-M., Hung, Y.-C., Chen, H.-G., & Ku, C.-Y. (2004). Transplanting the best practice for implementation of an ERP system: A structured inductive study of an international company. *Journal of Computer Information Systems*, *44*(4), 101–110.

Hung, Y., Huang, S., & Yen, D. C. (2004). A study on decision factors in adopting an online stock trading system by brokers in Taiwan. *Decision Support Systems*, *40*, 315–328. doi:10.1016/j.dss.2004.02.004

James, L. R., Demaree, R. G., & Wolf, G. (1993). R_{wg}: An assessment of within-group interrater agreement. *Journal of Applied Psychology*, *78*, 306–309.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York, NY: Holt, Rinehart, & Winston.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575.

Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of the *CVI*, *T*, $r_{wg(j)}$, and $r^*_{wg(j)}$ indexes. *Journal of Applied Psychology*, *84*, 640–647.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, *23*, 127–135.

Mak, D. C. S., Tsang, H. W. H., & Cheung, L. C. C. (2006). Job termination among individuals with severe mental illness participating in a supported employment program. *Psychiatry*, *69*, 239–248.

Maples, P., Franks, A., Ray, S., Stevens, A. B., & Wallace, L. S. (2010). Development and validation of a low-literacy chronic obstructive pulmonary disease knowledge questionnaire (COPD-Q). *Patient Education and Counseling*, *81*, 19–22. doi:10.1016/j.pec.2010.11.020

Mathews, H., Smith, S., Hussey, J., & Plack, M. M. (2010). Investigation of the preferred PT-PTA relationship in a 2:2 clinical education model. *Journal of Physical Therapy Education*, *24*(3), 50–61.

Moscoso, S., & Selgado, J. F. (2001). Psychometric properties of a structured interview to hire private security personnel. *Journal of Business and Psychology*, *16*, 51–59.

Namisango, E., Katabira, E., Karamagi, C., & Baguma, P. (2007). Validation of the Missoula-Vitas Quality-of-Life Index among patients with advanced AIDS in urban Kampala, Uganda. *Journal of Pain and Symptom Management*, *33*, 189–202. doi:10.1016/j.jpainsymman.2006.11.001

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? *Research in Nursing & Health*, *29*, 489–497.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*, 459–467.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson.

Rungtusanatham, M. (1998). Let's not overlook content validity. *Decision Line*, *29*, 10–13.

Saramma, P. P., & Thomas, S. V. (2010). Child rearing knowledge and practice scales for women with epilepsy. *Annals of Indian Academy of Neurology*, *13*, 171–179. doi:10.4103/0972-2327.70877

Sendjaya, S., Sarros, J. C., & Santora, J. C. (2008). Defining and measuring servant leadership behaviour in organizations. *Journal of Management Studies*, *45*, 402–424. doi:10.1111/j.1467-6486.2007.00761.x

Stausmire, J. M. (2011). Interdisciplinary development of an adult intubation procedural checklist. *Family Medicine*, *43*, 272–274.

Stelly, D. J. (2006, May). An explication of statistical significance testing applied to minimum content validity ratio (CVR) values. *2006 Society for Industrial and Organizational Psychology Conference*, Dallas, TX.

Tai, Y. (2011). Perceived value for customers in information sharing services. *Industrial Management & Data Systems*, *111*, 551–569. doi:10.1108/02635571111133542

Tai, Y., & Ho, C. (2010). Effects of information sharing on customer relationship intention. *Industrial Management & Data Systems*, *110*, 1385–1401. doi:10.1108/02635571011087446

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*, 358–376.

Tsang, H. W., & Wong, A. (2005). Development and validation of the Chinese version of Indiana Job Satisfaction Scale (CV-IJSS) for people with mental illness. *International Journal of Social Psychiatry*, *51*, 177–191. doi:10.1177/0020764005056766

Wallace, L. S., Gregory, H. B., Parham, J. S., & Baldridge, R. E. (2003). Development and content validation of family practice residency recruitment questionnaires. *Family Medicine*, *35*, 496–498.

Wei, C.-C. (2008). Evaluating the performance of an ERP system based on the knowledge of ERP implementation objectives. *International Journal of Advanced Manufacturing Technology*, *39*, 168–181. doi:10.1007/s00170-007-1189-3

Yu, S., Ng, C. S., Chang, S., Chang, I., & Yen, D. C. (2011). An ERP system performance assessment model development based on the balanced scorecard approach. *Information Systems Frontiers*, *13*, 429–450. doi:10.1007/s10796-009-9225-5

Zite, N. B., & Wallace, L. S. (2007). Development and validation of a medicaid postpartum tubal sterilization knowledge questionnaire. *Contraception*, *76*, 287–291. doi:10.1016/j.contraception.2007.06.012

Bios

F. Robert Wilson, PhD, is an emeritus professor of counseling of the University of Cincinnati with 35 years as a counselor educator. He completed doctoral studies at Michigan State University and post-graduate studies at the Cincinnati Gestalt Institute. His research interests include quantitative methods in counseling research, counselor education and supervision, and individual and group treatment of mental illness. He provides mental health counseling for indigent and homeless individuals with chronic mental illness.

Wei Pan, PhD, is an associate professor of quantitative research methodology at the University of Cincinnati. He received his doctorate in measurement and quantitative methods from Michigan State University in 2001 and his master's degree in mathematical statistics from Fuzhou University, China, in 1989. His research interests include causal inference, advanced statistical modeling, meta-analysis, and their applications in the social, behavioral, and health sciences.

Donald A. Schumsky, PhD, is an emeritus professor of psychology of the University of Cincinnati following a 45 year (42 at University of Cincinnati) career in teaching and research. His research interests include quantitative methods in psychological science, learning, motor skills, and cognition.